I/S: A JOURNAL OF LAW AND POLICY FOR THE INFORMATION SOCIETY

In Defense of Imprecision: Humanizing Big Data For Business Decision Making

ANGELA SHEN-HSIEH*

In the data analytics worlds of research, science, and academia, necessarily, there is much attention being given to precision, objectivity, de-biasing, and making sure the data is not misconstrued. Naturally, there are valid concerns about data lying. In particular, there are concerns about visualization because people "believe what they see," and visualization is, by definition, an abstract representation of data—with potentially dire implications if the representation is not quite right. Yet, researchers and those of us working with data, whether big or small, are also drawn to the use of visualization to communicate findings and insights, and enable exploration.

In the world of business analytics, where delivering quality information has been, and continues to be, a tremendous challenge, we have to break certain rules of scientific engagement with data in order to get data into the decision-making process in a useful way. Many practical considerations drive compromises and trade-offs to precision every day, but there are also benefits to shaping the data that's presented to our managers and decision makers in order to align them with a point of view of the business. This is actually one of the powerful capabilities of visualization. I believe we will see this tapped more and more as Big Data and the consumerization of data raise our expectations of value and usability. In this essay, I will walk you through some of my experience making data meaningful for businesses, and the place of data visualization in this effort.

I. BUSINESSES LOOK TO DATA FOR STORIES AND CONTEXT

Big Data in business is everything from web click streams and GPS "events," to the never-ending flow of information being generated from every organization's transactional systems each day. Even "small" data is pretty big and has been big long before there was the term 'Big Data'. Businesses are under tremendous pressure to use this to gain competitive advantage—they know they have much of the data they need to make better decisions, but still struggle to deliver actionable information. So, the pressure and potential of Big Data in the enterprise create a big incentive to figure out how to make it work.

Big Data has its own dynamics: there are technical challenges of course, but also quality concerns. For example, data from a single ping from a sensor on a vending machine outside a university building in Boston is neither interesting nor trustworthy. That single ping may be missing data or the sensor reading may be inaccurate, and there may be pings missing from a sequence. So, getting value out of Big Data is about the aggregation—the *forest* not the trees. Aggregating many of these pings, warts and all, can reveal many stories: what students substitute for breakfast in the morning when it's cold out or crave after class in the afternoon. Monitor this data over time and combine it with data from other machines across the country, and you paint a picture of the way eating habits trend and migrate nationally. In this way, imperfect data—facts almost useless at the leaf level—can be formed into a clear picture at an aggregate level. However, this can be counterintuitive for many data scientists accustomed to ensuring a solid foundation of data quality before building up their stacks of statistical correlations and logical conclusions.

But business data is most often provided as context for a human thought process—a lower relevance bar to hurdle. Definitive decisions, correlations, or causality would be nice, but are rarely found or trusted. Instead, we use data to sketch out a narrative of the situation at hand and engage human intuition and experience to reach conclusions. So, to get any of this information to fulfill its promise—to be successful in changing behavior and outcomes, to be impactful to lives and livelihoods—the information has to connect with hearts and minds. Accomplishing this is a soft art, for sure, but we've also had to shift the focus from perfectly capturing every transaction (and not losing anything!) to delivering as much information as possible, perfect or not, in ways that provide context and narrative, and to injecting data with the right perspectives to incite action and decision-making. This can mean using approximations or proxies for missing or incomplete context, distorting and biasing the presentation of the

data in order to promote certain factors or prioritize specific conclusions, emphasizing a specific analysis path to align decision-making with organizational objectives, and even truncating data and obscuring low priority information. This is a "ready-fire-aim" approach -- not at all an exact science. It's an approach critical to keeping pace with rapidly evolving business dynamics, with precision and objective truth sacrificed for speed.

II. LEARNING FROM THE LAST FEW DECADES OF BUSINESS ANALYTICS

I come from the messy world of business data. For decades, this world's primary focus was capturing transactional data for, primarily, accounting and regulatory reporting. Tremendous amounts of money, effort, and technology went into accurately capturing and storing every purchase order, sale, shipment, invoice payment, item of inventory, etc. For decades, software and technology vendors have rallied customers on the marketing mantra of "the single version of the truth." Obviously, accurate reporting of every cent coming in and going out is critical for any business, and in particular public companies. Somewhere along the way, businesses realized that this data should be used to drive better decisions. And the opportunities to drive more competitive and successful business outcomes through data analysis seem so tantalizingly possible. What if we knew exactly how our customers would act by analyzing every interaction they have with our brand and company? What if we could optimize every asset we had by monitoring sensors on every piece of equipment to schedule maintenance before a failure, maximize scheduling and logistics, and reduce idle time?

However, the data was captured and structured for fiscal reporting and not necessarily for decision support. This reality, in turn, drove the next several decades of IT investment–technologies to not only capture broad and diverse data in businesses – for example, about customers (known as *customer relationship management* or "CRM" systems), or plans and operations (known as *enterprise resource planning* or "ERP" systems – but to model and deliver it to knowledge workers in more meaningful and actionable ways through data storage and access technologies and reporting and presentation tools.

Enterprise data is messy because data is really just a proxy: Data doesn't actually start out as data. Data starts out as a category of information that everyone has agreed is valuable to capture and archive. But in order to do that, we have to take that transaction or event, plan, interaction, or idea and strip it down into component facts that we can capture and place into cells of a database. We essentially

remove much of the context of the information in order to be able to store and retrieve it. When we do then try to retrieve it and put it into some kind of a narrative structure for human consumption, the pieces do not necessarily go back together or line up neatly. Sorting out this mess is then the purpose of much of these business technologies and especially of data visualization—putting back context and connective tissue that enables decision-makers to take the data and move it up the chain to information, knowledge, and aspiringly—to wisdom. But getting the data up to the user ("to the glass" of your computer's screen) is only half of the equation. The context of how the data will be used, how it fills gaps in a decision-making process, how it helps to address a problem – that's the other half. To get data presentation and data-driven decision-making seamlessly to meet somewhere in the middle requires give-and-take on both sides. It's in that compromise where precision is sacrificed.

III. WHAT'S DIFFERENT ABOUT THE USE OF DATA IN BUSINESS?

In business analytics, complete objectivity is not the singular goal, because data in business is about decision-making. Making decisions every day in a business is naturally filled with subjectivity and fraught with 'fuzzy' logic. It's not the same as defending a position or proving a conclusion from a set of research. The end goals are different.

A. *Important decisions are "people" decisions.*

Critical and strategic business decisions are rarely driven by an algorithm, as in, what ad to serve to a web page. In business, the high-value decisions—what to invest in, whom to hire—are made by people. In business, we have to get data into consumable forms for people so that it supports the actions and decisions they need to make. We use things like reports, dashboards, charts, and data visualization to do this, but these tools are not designed to give definitive answers to their users. If they did, we wouldn't need charts and visualization; we'd just tell decision makers what to do or take human judgment out of the loop by automating the action. What presentations of data need to do is convey information for human processing and enable the power of modern computing to seamlessly work in support of the power of the human brain. The human brain looks for patterns and invokes experience, intuition, and "gut feel." In business, we pay those decision makers a lot of money to have those things and be able to apply them—but these capabilities are subjective and very hard to quantify. So a lot of what we are really trying to do with data in

business is align it with the not-so-linear, not-so-logical, not-necessarily-defensible thought processes and inner workings of people. A number alone is hard-pressed to establish the trust or orientation needed for better decisions. High-value decisions in business are supported by data, but they are rarely derived solely from data. For instance, it seems unlikely that there was statistical evidence that consumers would embrace a 10-inch phone when Apple decided to invest in the iPad, right?

B. *We are looking for business relevance.*

What is statistically relevant may not be business-relevant or may not point to an obvious decision. The causal relationship may not be apparent, and it may not be clear what to do based on that information, so it's often not worth the effort to achieve a precise degree of confidence. A simple algorithm might tell me that there is an increase in e-purchase of yellow cars in Texas, but, as a manufacturer, that doesn't mean I'm going to increase the output of yellow cars and start shipping them south. To understand whether this has any business relevance that I need to act on, I'm going to want to explore what's going on in Texas—is this a local phenomenon to one or a handful of dealers? What was the inventory of yellow cars before this trend, and are the sales related to some promotion to get rid of those cars? Could these purchases be related to something else in the environment—the rise of a popular yellow brand or sports team? Or are these yellow car orders for taxicabs? As a person responsible for tracking these trends and making production decisions, I know there can be external factors that drive oddities in the data. I also know that yellow is the hardest color of cars to sell, and I do not want to end up with an inventory heavily loaded with cars I cannot sell.

C. *Data is delivered to align people.*

In the enterprise, we go to tremendous trouble to disseminate data. Modern information technology – the first 30 or 40 years of it – is about capturing data and getting it to people. The way people consume data is designed to support individual decision-making. But, implicit in this connection is that data and the way it's delivered aligns a firm around its organizational strategy and the key objectives that support the success of that strategy. For example, businesses use KPIs (Key Performance Indicators) to measure how they are doing. These KPIs evolve and change constantly. A business might develop a

strategy to take advantage of a market opportunity or to improve the way they are doing something operationally. They want to be able to execute on that. So the way they align a lot of teams and people is to create specific metrics generated from data to measure performance against these goals. These KPIs and the context in which they are presented are inherently biased. For example, a company might identify an operational weakness and design a strategy focused on improving their speed-to-market in a certain way. Certain aspects—like improving planning or resource allocation or having timely functional bandwidth and hiring –are being prioritized above others. The business will want to monitor these factors closely and weigh decisions on these metrics above others like budgets and efficiencies. The data is modeled and presented to exploit the specific parameters of interest—with the intent to get everyone focused on optimizing this circumstance. So, the way data is presented in reports, visualizations, and data-driven analytic applications needs to convey the organization's perspective on the information. In this way, those of us working in business analytics actually strive for data with a point of view. We want the subjectivity of the business because the whole point in business is better decision making—and 'better' is subjective, company-specific, and often a moving target.

IV. AN EXAMPLE OF SOME OF THE DATA CHALLENGES

What if, as a manufacturing company, I am trying to understand the quality of service we are providing to our customers and want to look at the timeliness of shipments. This sounds pretty straightforward. But to track the shipments to even one customer is actually very complex. For instance, we would need to follow the data trail from the system that takes orders to the systems that fulfill the orders and ship them. Then, we would need to be able to get the data from our external freight and shipping vendors to know when goods were actually delivered. If there were problems or complaints, we would need to cross all this with the information from our customer service systems—much of which may be unstructured data, like phone call transcripts, requiring text mining to confirm a positive or negative nature of the call or severity of the problem—to turn that "info" into something structured that can be similarly analyzed.

At the heart of this project is being able to have every transaction traceable to a unique customer. This seemingly simple requirement can be almost unattainable for large, multinational organizations. Is Acme Ltd in one system the same as Acme Limited in another? Should sales to Acme-ABC Manufacturing, a joint venture between Acme and

ABC be attributed to the Acme customer column or the ABC customer column? The sales team in China does it differently from the U.S. team. Then, most large companies are continually acquiring new companies, which may share some of the same customers, but not the same definitions or systems. What do I do? Perhaps I look at service quality only for our legacy businesses and the new businesses will have to wait until the harmonization and systems consolidation process is complete (if ever). Already, I am making big compromises on the quality of the data and analysis.

So, in order to know if we are improving, we want to get to some kind of output that gives us the general idea of how we are doing—just okay, or it is really ugly— where, with what type of customers, with what size customers, in what areas of the business, in what areas of the world, and on and on. A "score" that we can aggregate up and across would make it possible to compare apples to oranges.

And this quality score has to account for a somewhat subjective experience of the customer, and it is highly dependent upon expectations, i.e., "I was told the items were out-of-stock but received them in four days" (happy customer) vs. "I paid for two-day shipping but didn't receive the items until after four days (unhappy customer)." How do we factor customer expectation into our quality scoring model? This is an important factor in the analysis, but I am going to have to be satisfied with some sort of proxy for meets/does not meet the expectation (perhaps a combination of whether my shipping vendor reported an issue or a manual note was made by an account manager, or possibly a sentiment analysis of support call logs), and I am going to have to understand that an unknown percentage of these shipments is likely to be tagged incorrectly or not tagged at all.

Then, do we score at the order level, i.e., this order gets an A, this one a D? But there may be many shipments involved in filling one order, coming from different locations delivered by different carriers. So the shipment level would be more precise, but much more computationally intensive. However, if we do not analyze the data at the shipment level, we won't be able to identify warehouse issues or issues with certain delivery companies or routes. Should we just aggregate everything to a customer level-- and in that case, is this a useful exercise at all?

This is the thought process business analysts go through every day, weighing the many trade-offs among complexity, flexibility, effort, and value. And reflected in these choices are very much points of view based on the priorities and interests of the business at hand. If the business has few levers to pull on the fulfillment and delivery side, but feels confident in its communications with its customers and sees

better expectation-setting as an opportunity to strengthen the firm's relationship with and knowledge of its customers, then a more general sense of customer satisfaction is enough to know whom to target with more hands-on attention. There may be important insights to be found in the shipment tracking data, but, if they are not actionable, we are not going to show them. Interestingly, in this way, the capacity of the data to provide analysis actually drives the strategy and types of actions available to the business. If I cannot actually measure the quality of each delivery, how can I pressure my shippers to better performance?

V. VISUALIZATION IS ABOUT MAKING DATA CONSUMABLE, BUT PRECISION CAN IMPEDE COMPREHENSION AND DECISION MAKING

For professionals charged with curating and presenting data for use in business settings, the dynamics and the uncertainties I have described yield four critical lessons.

A. *The exact number is not always important.*

For decision-making, it is rarely the exact number that's important—it is the relationship of that number to a larger context. If I am looking at performance numbers, such as revenues or expenses, from different regions, the numbers alone are not meaningful unless I know the context surrounding what I was expecting: What is my target or goal? How did this year's numbers compare to last year's numbers? If a predictive algorithm gives me a set of probabilities for, say, a set of options, it is not enough to be actionable in a business context to know that option A has a probability of 64%, and option B has a probability of 58%. A manager or decision-maker is not going to invest dollars, people, or time until he or she understands much more about the dynamics and drivers of the situation and of the predictive model. The number alone is not context, and other information must be provided and presented appropriately.

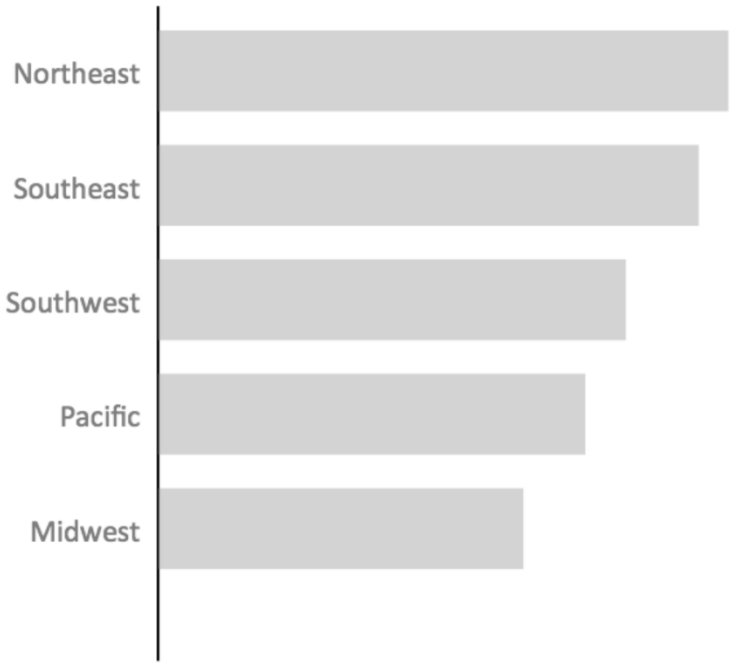


Figure 1: Little context = not actionable. "Is this a good or bad story?"

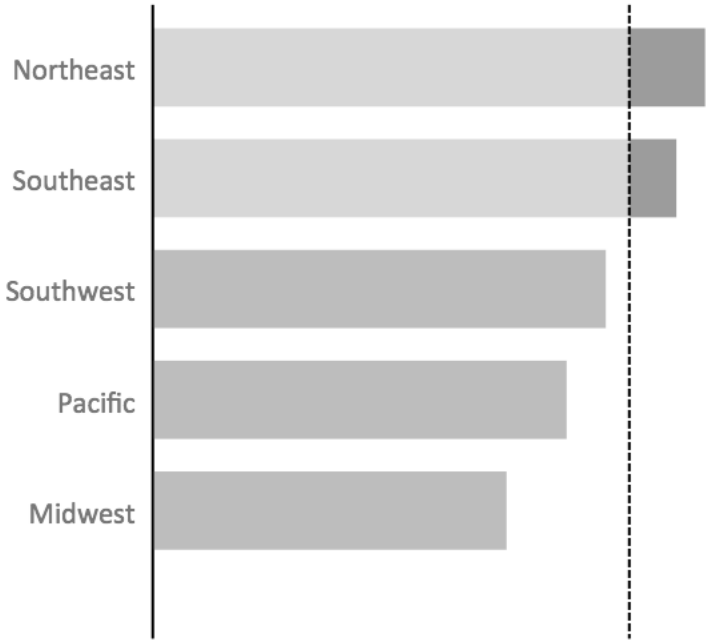


Figure 2 | A goal line and color provides context. "Three regions are behind."

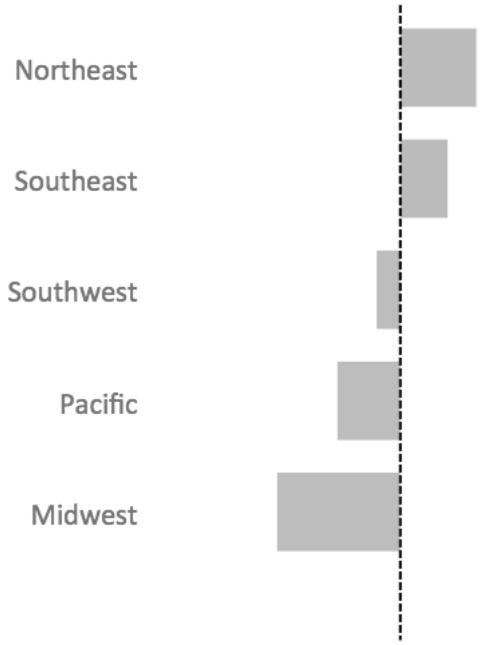


Figure 3 | Here the emphasis is on comparing the regions. "Midwest is twice as far behind as Pacific"

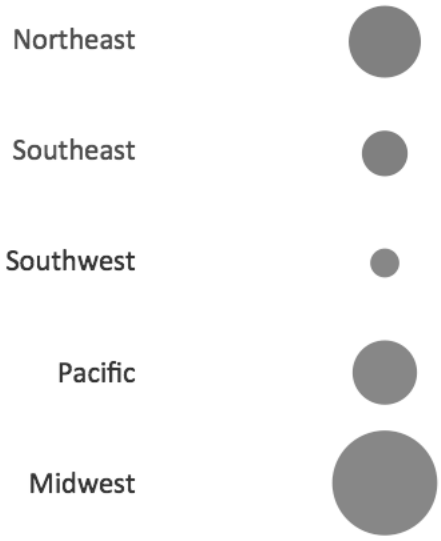


Figure 4: Less precise to enable fast identification of areas needing attention "What's going on in the Midwest?!"

B. *Visualization is a metaphor.*

In business, the way we deliver data for human consumption is through visualization and visualization works through metaphor. Metaphor is, by its nature, an approximation. Much of the emphasis in visualization has been about the proper use of well-known chart and graph archetypes and the creation of new visualization types. But a good visualization in business is really a visual metaphor for a business circumstance and the way we think about our business. Simply displaying the businesses data correctly is not enough when situations are complex. There are multiple dimensions and criteria to consider. More than likely, we are going to have to act without complete or perfect data. Richer visual metaphors can evoke the "gut feel" and creativity necessary to assess the significance of the situation at hand.

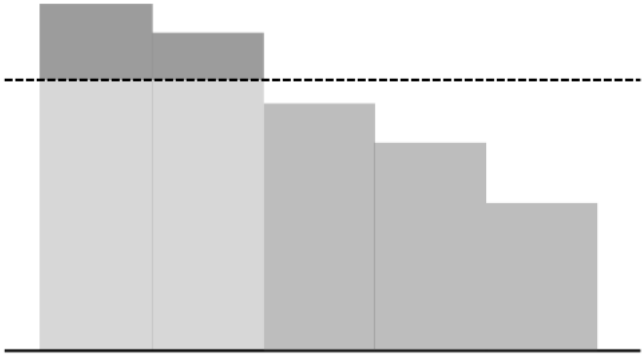


Figure 5 | Same visual metaphor, flipped on its side, becomes a trend over time.

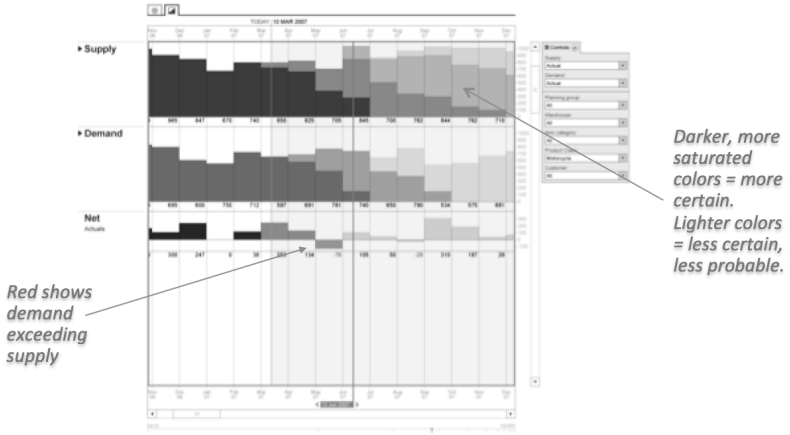
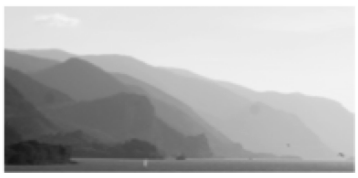


Figure 6 | The visual metaphor becomes richer to tell a story of past and future supply and demand.



http://www.flickr.com/photos/wwworks/2712147416/sizes/l/in/photostream/

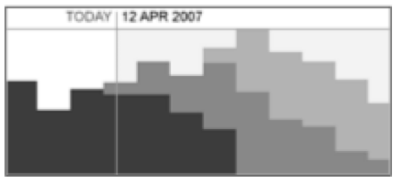


Figure 7 | The visual metaphor depicts lighter elements as "off in the distance" and "less clear/certain" by mimicking the way the eye perceives things in nature.

C. *All data does not have to be treated equally.*

It seems inarguable that for good data analysis, data must be looked at objectively. Each point of data could have relevance, so it is equally important and should be considered in the same way. Our results must be free of bias. Visualization, with its roots in the sciences, has followed this accepted wisdom. However, when analyzing business data, we readily discount and exclude swaths of information or prioritize certain factors because of ease-of-use and business relevance. In visualization, we might intentionally promote or demote data because of its relevance to the decision-makers' thought process. If I am looking at my sales globally, many of the countries across the world will likely have small amounts of revenue and in aggregate make up only a small percentage of my total business. Others we may exclude because they may be in markets where we are divesting or where we do not face heavy competition. These would not be good markers for performance or trending. Similarly, showing the data for 150 countries, while thorough and likely the correct response to the query made, is going to generate a chart or table that either reduces the size of elements in order to fit in a single view (as in a scatterplot or treemap) or requires the viewer to scroll (as in a bar chart or a table) either of which is likely to obscure or distract from important information. In this way, objectivity can sometimes be an impediment to human analysis.

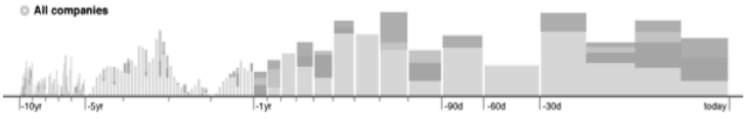


Figure 8 | Time in the distant past (x axis) is compressed to focus the analysis on recent events.

D. *Providing a point of view on data can drive changes in behavior and outcomes.*

Data analysis these days is often likened to storytelling. Storytelling has a point of view, establishes a narrative structure, and is intended to evoke feelings and emotion. Stories are one of the ways we better understand and remember things. Stories are used in business every day to explain what's happening and create the right

attitude to move forward. A start-up might motivate its team by portraying its competitive situation as a David meets Goliath epic battle, or a manager may frame her objective as a before-and-after transformation story. Data about your personal health likely contains information that you should act on, whether it's losing a few pounds or eating less red meat. But having the data–a number on the scale or your cholesterol score–doesn't mean you act accordingly. Whether it's describing a negative trajectory and some very undesirable possible outcomes – e.g., diabetes and daily insulin shots, heart disease and not living to see your grandchildren – or simply connecting some dots (when you order soda at a restaurant, you may be doubling the calories of that meal), a narrative helps people to absorb the information in a more meaningful way. It provides them more context for the decisions they make--for example, at mealtime-- every day. In the same way, access to the facts about the health of a business doesn't ensure the obvious actions are taken.

When we are trying to support human decision making, there is more at play than just cognitive comprehension; there is the wild card of human emotion, which is now becoming understood as an important factor in decision-making. If we do not like the message in the data, or it disagrees with our thinking, we may question the measurement process, or the relevance to the current situation-- or ignore the findings altogether. If the data and the presentation does not engage us -- we do not like the “messenger” -- we may not discern the story there or know how to act on it.

An example that comes up a lot involves showing data from other teams: for instance, whether product development teams should be able to see the progress and performance of other teams in a corporate dashboard. Some organizations want teams to see other teams’ progress because they believe transparency promotes positive things like keeping the data up-to-date and healthy competition. Other organizations worry that the visibility of other teams' data may foster bad behavior, like misreporting, or hurt morale and dampen motivation. The design of the dashboard can impact the teams' emotional reaction to the information and mediate these concerns. A light, friendly look and feel can change the impression. Taking an even hand to show good and bad performance -- for example, avoiding bright red and green stoplights and giving more visual emphasis on progress within range than to outliers -- can soften the message in the data. In this way, people and teams lagging (in the "red") do not feel they are being called out on the carpet publicly. The choice of the data creates its own narrative, and in its presentation, there are also many choices to best enable these users to accept and understand the

findings in the data and feel compelled to act on them appropriately.

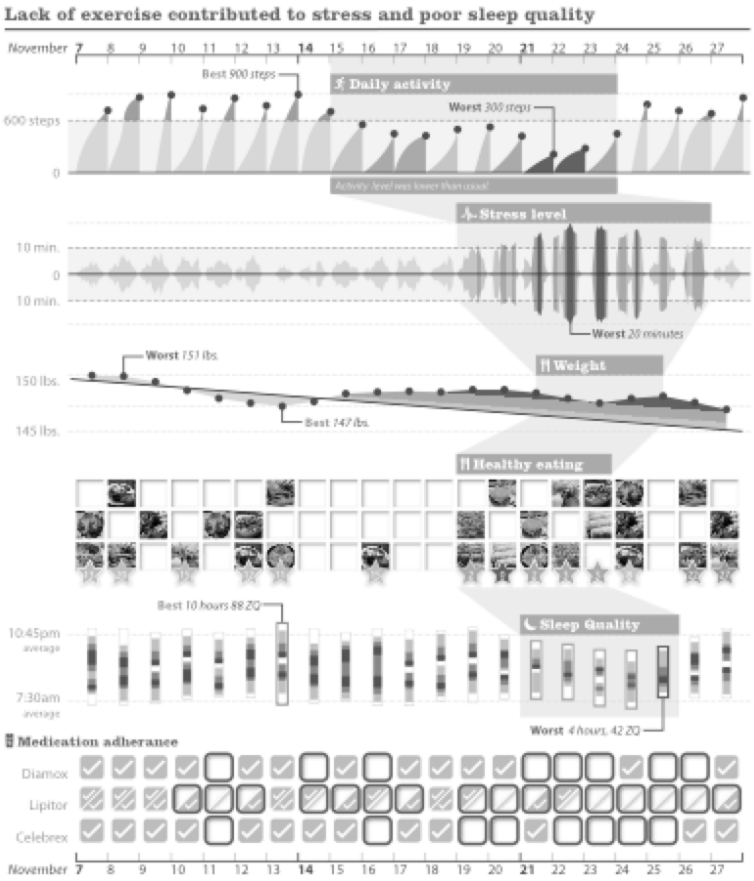


Figure 9 | Data visualization designed to engage people with their personal health data. Yellow areas highlighted the connection between Low Activity, High Stress, Poor Eating and Sleep Quality.

Data has always had an aura of "truth": it's a *fact*, it must be *right*, there's something *precious* about each and every point so painstakingly captured and archived. But the dynamics of Big Data and broad consumer access to data are changing that perception and it's not a bad thing. There's no one-size-fits-all data approach. And the human experience of the users, their objectives, and those of their organizations is making data analysis and visualization as much an art as a science.